Dated 03-11-16

DRAFT Technical Requirements Document

APEX 2020

LA-UR-15-28541
SAND2016-2034 O

Lawrence Berkeley National Laboratories is operated by the University of California for the U.S. Department of Energy under contract NO. DE-AC02-05CH11231.

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. LA-UR-15-28541 Approved for public release; distribution is unlimited.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND2016-2034 O.

Dated 03-11-16

APEX 2020:
Draft Technical Specifications

1 Introduction

Los Alamos National Security, LLC (LANS), in furtherance of its participation in the Alliance for Computing at Extreme Scale (ACES), a collaboration between Los Alamos National Laboratory and Sandia National Laboratories; and in coordination with the Regents of the University of California, which operates the National Energy Research Scientific Computing (NERSC) Center residing within the Lawrence Berkeley National Laboratory (LBNL), is releasing a joint Request for Proposal (RFP) under the Alliance for application Performance at EXtreme scale (APEX). The APEX 2020 RFP will be for two next generation systems, Crossroads and NERSC-9, to be delivered in the 2020 time frame. The selected vendor will be responsible for delivering and installing the Crossroads and NERSC-9 systems at their respective locations. While our preferred option is to award both the Crossroads and NERSC-9 subcontracts to a single Offeror, awards may be made to separate Offerors. Awards will be made separately by ACES and NERSC. In total there will be four subcontracts, two Non-Recurring Engineering (Section 4) and two build (an NRE and build contract per entity). The technical specifications in this document describe joint requirements wherever possible. Where requirements for Crossroads and NERSC-9 differ it will be made clear in this document.

Crossroads and NERSC-9 each have maximum funding limits over their system lives, to include all design and development, site preparation, maintenance, support and analysts. Total ownership costs will be considered in system selection. The Offeror must respond with a configuration and price for both systems.

Application performance and workflow efficiency are essential to these procurements. Success will be defined as meeting our 2020 mission needs while at the same time serving as a pre-exascale platform that enables our applications to begin to evolve using yet to be defined next generation programming models. The advanced technology aspects of the APEX platforms will be pursued both by fielding first of a kind technologies on the path to exascale as part of system build and by selecting and participating in strategic NRE projects with the Offeror and applicable technology providers.

1.1 Crossroads

The Department of Energy (DOE) National Nuclear Security Administration (NNSA) Advanced Simulation and Computing (ASC) Program requires a computing system be deployed in 2020 to support the Stockpile Stewardship Program. In the 2020 timeframe, Trinity, the first ASC Advanced Technology System (ATS-1), will be nearing the end of its useful lifetime. Crossroads, the proposed ATS-3 platform, provides a replacement, tri-lab computing resource for existing simulation codes and provides a larger resource for ever-increasing computing requirements to support the weapons program. The Crossroads system, to be sited at Los Alamos, NM, is projected to provide a large portion of the ATS resources for the NNSA ASC tri-lab simulation community: Los Alamos National Laboratory (LANL), Sandia National Laboratories (SNL), and Lawrence Livermore National Laboratory (LLNL), during the 2021-2025 timeframe.

In order to fulfill its mission, the NNSA Stockpile Stewardship Program requires higher performance computational resources than are currently available within the Nuclear Security Enterprise (NSE). These capabilities are required for supporting stockpile stewardship certification and assessments to ensure that the nation's nuclear stockpile is safe, reliable, and secure.

The ASC Program is faced with significant challenges by the ongoing technology revolution. It must continue to meet the mission needs of the current applications but also adapt to radical change in technology in order to continue running the most demanding applications in the future. The ASC Program recognizes that the simulation environment of the future will be transformed with new computing architectures and new programming models that will take advantage of the new architectures. Within this context, ASC recognizes that ASC applications must begin the transition to the new simulation environment or they may become obsolete as a result of not leveraging technology driven by market trends. With this challenge of technology change, it is a major programmatic driver to provide an architecture that keeps ASC moving forward and allows applications to fully explore and exploit upcoming technologies, in addition to meeting NNSA Defense Programs' mission needs. It is possible that major modifications to the ASC simulation tools will be required in order to take full advantage of the new technology. However, codes running on NNSA Advanced Technology Systems (Trinity and Sierra) in the 2019 timeframe are expected to run on Crossroads. In some cases new applications also may need to be developed. Crossroads is expected to help technology development for the ASC Program to meet the requirements of future platforms with greater computational performance or capability. Crossroads will serve as a technology path for future ASC systems in the next decade.

To directly support the ASC Roadmap, which states that "work in this timeframe will establish a strong technological foundation to build toward exascale computing environments, which predictive capability may demand," it is critical for the ASC Program to both explore the rapidly changing technology of future systems and to provide platforms with higher performance and more memory capacity for predictive capability. Therefore, a design goal of Crossroads is to achieve a balance between usability of current NNSA ASC simulation codes and adaptation to new computing technologies.

1.2 NERSC-9

The DOE Office of Science (SC) requires a high performance production computing system in the 2020 timeframe to provide a significant upgrade to the current computational and data capabilities that support the basic and applied research programs that help accomplish the mission of DOE SC.

The system also needs to be a platform that will provide a firm foundation for future exascale systems in 2023 and beyond, a need that is called out in the DOE's Strategic Plan 2014-2018, that calls out for "advanced scientific computing to analyze, model, simulate and predict complex phenomena, including the scientific potential that exascale simulation and data will provide in the future."

NERSC Center supports nearly 6000 users and about 600 different application codes from a broad range of science disciplines covering all six program offices in SC. The scientific goals that are well summarized in the 2012-2014 series of requirements reviews commissioned by the Advanced Scientific Computing Research (ASCR) office that brought together application scientists, computer scientists, applied mathematicians, DOE program managers and NERSC personnel. The 2012-2014 requirements reviews indicated that compute-intensive research and research that attempts scientific discovery through the analysis of experimental and observational data both have a clear need for major increases in computational capability and capacity in the 2017 timeframe and beyond. In addition, several science areas also have a burgeoning need for HPC resources that satisfy an increased compute workload and provide strong support for data-centric workflows and real-time observational science. More details about the DOE SC application requirements are in the reviews located at:

http://www.nersc.gov/science/hpc-requirements-reviews/

NERSC has already begun transitioning the SC user base to energy efficient architectures, with the procurement of the NERSC-8 "Cori" system. In the 2020 time frame, NERSC also expects a need to address early exascale hardware and software technologies, including the areas of processor technology, memory hierarchies, networking technology, and programming models.

The NERSC-9 system is expected to run for 4-6 years and will be housed in the Wang Hall (Building 59) at LBNL that currently houses the "Cori" system and other resources that NERSC supports. The system must integrate into the NERSC environment and provide high bandwidth access to existing data stored by continuing research projects. For more information about NERSC and the current systems, environment, and support provided for our users, see http://www.nersc.gov.

1.3 Schedule

The following is the tentative schedule for the Crossroads and NERSC-9 systems.

Table 1 Crossroads/NERSC-9 Schedule

	Crossroads and NERSC-9
RFP Released	Q2CY16
Subcontracts (NRE/Build) Awarded	Q4CY16
On-site System Delivery Begins	Q2CY20
On-site System Delivery Complete	Q3CY20
Acceptance Complete	Q1CY21

2 Mandatory Requirements

An Offeror shall address all Mandatory Requirements and its proposal shall demonstrate how it meets or exceeds each one. A proposal will be deemed non-responsive/unacceptable, will be rejected, and will not be considered further if each and every one of the following Mandatory Requirements is not met.

The Offeror's proposal shall be structured such that each Mandatory Requirement and Technical Design Requirement included in this document is listed verbatim in the Offeror's response followed by the Offeror's response to that requirement. The proposal shall clearly describe the role of any subcontractor(s) and the technology or technologies, both hardware and software, and value add they provide where appropriate. The Offeror shall respond with a single proposal that contains distinct sections showing how and where their proposed Crossroads and NERSC-9 systems differ. Alternatives to hardware, software, and/or architectural solutions may also be included in the Offeror's single proposal.

2.1.1 The Offeror shall provide a detailed full platform architectural description of both the Crossroads and NERSC-9 systems, including diagrams and text describing the following details as they pertain to the Offeror's platform architecture(s):

- Component architecture – details of all processor(s), memory technologies, storage technologies, network interconnect(s) and any other applicable components.
- Node architecture(s) – details of how components are combined into the node architecture(s). Details shall include bandwidth and latency specifications (or projections) between components.
- Board and/or blade architecture(s) – details of how the node architecture(s) is integrated at the board and/or blade level. Details should include all inter-node and inter-board/blade communication paths and any additional board/blade level components.

- Rack and/or cabinet architecture(s) – details of how board and/or blades are organized and integrated into racks and/or cabinets. Details should include all inter rack/cabinet communication paths and any additional rack/cabinet level components.
- Platform storage – details of how storage is integrated with the proposed system, including a storage architectural diagram.
- Platform architecture – details of how rack or cabinets are combined to produce platform architecture, including the high-speed interconnects and network topologies (if multiple) and storage.
- Proposed floor plan – including details of the physical footprint of the proposed system and all of the supporting components.

2.1.2 The Offeror shall provide a detailed description of the software eco-system, including a high-level software architectural diagram including the provenance of the software component, for example open source or proprietary and support mechanism for each.

2.1.3 The Offeror shall describe how the proposed system does or does not fit into the Offeror's long-term product roadmap and a potential follow-on platform acquisition in the 2025 and beyond timeframe.

3 Target Design Requirements

This section contains detailed system design targets and performance features. It is desirable that the Offeror's design meets or exceeds all the features and performance metrics outlined in this section. If a Target Design Requirement cannot be met, it is desirable that the Offeror provide a development and deployment plan, including a schedule, to satisfy the requirement.

Proposals that do not address the Target Design Requirements in a materially responsive manner will be downgraded. The Offeror shall also propose any hardware and/or software architectural features that will provide improvements for any aspect of the proposed system.

3.1 Scalability

The scale of the platform necessary to meet the needs of the application requirements of the APEX laboratories adds significant challenges. The Offeror shall propose a system that enables application performance up to the full scale of the platform. Additionally, the system proposed should provide functionality that assists users in obtaining performance at up to full scale. Scalability features, both hardware and software, that benefit both current and future programming models are essential.

3.1.1 The proposed system shall support running jobs up to and including the full scale of the system.

3.1.2 The proposed system shall support launching an application at full system scale in less than 30 seconds. The Offeror shall describe factors (such as executable size) that could potentially affect application launch time.

3.1.3 The Offeror shall describe how applications launch scales with the number of concurrent launch requests (pers second) and scale of each launch request (resources requested, such as the number of scheduleable units etc.), including information such as:

- All system-level and node-level overhead in the process startup including how overhead scales with node count for parallel applications, or how overhead scales with the application count for large numbers of serial applications.
- Any limitations for processes on compute nodes from interfacing with an external work-flow manager, external database or message queue system.

3.1.4 The proposed system shall support thousands of concurrent users and more than 20,000 concurrent batch jobs. The system shall allow a mix of application or user identity wherein at least a subset of nodes can run multiple independent applications from multiple users. The Offeror shall describe details, including limitations of their proposed support for this requirement.

3.1.5 The Offeror shall describe all areas of the proposed system in which node-level resource usage (hardware and software) increases as a job scales up (node, core or thread count).

3.1.6 The proposed system shall utilize an optimized job placement algorithm to reduce job runtime, lower variability, minimize latency, etc. The Offeror shall describe in detail how the algorithm is optimized to the system architecture.

3.1.7 The Offeror shall provide an application programming interface to allow applications access to the physical to logical mapping information of the job's node allocation – including a mapping between MPI ranks and network topology coordinates, and core, node and rack identifiers.

3.1.8 The Offeror shall describe how the system software solution provides a low jitter environment for applications and shall provide an estimate of a compute node operating system's noise profile, both while idle and while running a non-trivial MPI application. If core specialization is used, describe the system software activity that remains on the application cores.

3.1.9 The system shall provide correct numerical results and consistent runtimes (i.e. wall clock time) that do not vary more than 3% from run to run in dedicated mode and 5% in production mode. The Offeror shall describe strategies for minimizing runtime variability.

3.1.10 The proposed system's high speed interconnect shall support a high messaging bandwidth, high injection rate, low latency, high throughput, and independent progress. The Offeror shall describe:

- The system interconnect in detail, including any mechanisms for adapting to heavy loads or inoperable links, as well as a description of how different types of failures will be addressed.
- How the interface will allow all cores in the system to simultaneously communicate synchronously or asynchronously with the high speed interconnect.
- How the interconnect will enable low-latency communication for one- and two-sided paradigms.

3.1.11 The Offerer shall describe how both hardware and software components of the interconnect support effective computation and communication overlap for both point-to-point operations and collective operations (i.e., the ability of the interconnect subsystem to progress outstanding communication requests in the background of the main computation thread).

3.1.12 The Offeror shall report or project the node injection/ejection bandwidth.

3.1.13 The Offeror shall report or project the bit error rate of the interconnect in terms of time period between errors that interrupt a job running at the full scale of the proposed platform.

3.1.14 The Offeror shall describe how the interconnect will provide Quality of Service (QoS) capabilities (e.g., in the form of virtual channels or other sub-system QoS capabilities), including but not limited to:

- An explanation of how these capabilities can be used to prevent core communication traffic from interfering with other classes of communication, such as debugging and performance tools or with I/O traffic.
- An explanation of how these capabilities allow efficient adaptive routing as well as a capability to prevent traffic from different applications interfering with each other (either through QoS capabilities or appropriate job partitioning).
- An explanation of any sub-system QoS capabilities (e.g. storage system QoS features).

3.1.15 The Offeror shall describe specialized hardware or software features of the system that accelerate workflows or components of workflows such as data analysis or visualization, and describe any limits their scalability on the proposed system. The hardware shall be on the same high speed network as the main compute resources and shall have equal access to other compute resources (e.g. file systems and storage). It is desirable that the hardware have the same node level architecture as the main compute resources.

3.2 System Software and Runtime

The Offeror shall propose a well-integrated and supported system software environment. The overall imperative is to provide users with a productive, high-performing, reliable, and scalable system software environment that enables efficient use of the full capability of the proposed system.

3.2.1 The proposed system shall include a full-featured Linux operating system environment on all user visible service partitions (e.g., front-end nodes, service nodes, I/O nodes). The Offeror shall describe the proposed full-featured Linux operating system environment.

3.2.2 The proposed system shall include an optimized compute partition operating system that provides an efficient execution environment for applications running up to full-system scale. The Offeror shall describe any HPC relevant optimizations made to the compute partition operating system.

3.2.3 The Offeror shall describe the security capabilities of the operating systems proposed in technical requirements 3.2.1 and 3.2.2.

3.2.4 The proposed system shall provide efficient support for dynamic shared libraries, both at job load time and during runtime. The Offeror shall describe how applications using shared libraries will execute at full system scale with minimal performance overhead compared to statically linked applications.

3.2.5 The Offeror shall provide resource management functionality, including job migration, backfill, targeting of specified resources (e.g., platform storage), advance and persistent reservations, job preemption, job accounting, architecture-aware job placement, power management, job dependencies (e.g., workload management), and resilience management. The Offeror may propose multiple options for a vendor-supported resource manager.

3.2.6 The proposed system shall support jobs consisting of multiple individual applications running simultaneously (inter- or intra-node) and cooperating as part of an overall multi-component application (e.g., a job that couples a simulation application to an analysis application). The Offeror shall describe in detail how this will be supported by the system software infrastructure (e.g., user interfaces, security model, and inter-application communication).

3.2.7 The Offeror shall describe a mechanism that will allow users to provide containerized software images without requiring privileged access to the system or allowing a user to escalate privilege. The startup time for launching a parallel application in a containerized software image at full system scale should not greatly exceed the startup time for launching a parallel application in the vendor-provided image.

3.2.8 The Offeror shall describe a mechanism for dynamically configuring external IPv4/IPv6 connectivity to and from compute nodes, enabling special connectivity paths for subsets of nodes on a per-batch-job basis, and allowing fully routable interactions with external services.

3.2.9 The Offeror shall provide access to source code, and necessary build environment, for all software except for firmware, compilers, and third party products. The Offeror shall provide updates of source code, and any necessary build environment, for all software over the life of the subcontract.

3.3 Software Tools and Programming Environment

The primary programming models used in production applications in this time frame are the Message Passing Interface (MPI), for inter-node communication, and OpenMP, for fine-grained on-node parallelism. While MPI+OpenMP will be the majority of the workload, the APEX laboratories expect some new applications to exercise emerging asynchronous programming models. System support that would accelerate these programming models/runtimes and benefit MPI+OpenMP is desirable.

3.3.1 The Offeror shall provide an implementation of the MPI version 3.1 (or then current) standard specification. The Offeror shall provide a detailed description of the MPI implementation (including specification version) and support for features such as accelerated collectives, and shall describe any limitations relative to the MPI standard.

3.3.2 The Offeror shall describe at what parallel granularity the system can be utilized by MPI-only applications.

3.3.3 The Offeror shall provide optimized implementations of collective operations utilizing both inter-node and intra-node features where appropriate, including MPI_Barrier, MPI_Allreduce, MPI_Reduce, MPI_Allgather, and MPI_Gather.

3.3.4 The Offeror shall describe the network transport layer of the proposed system including support for OpenUCX, Portals, libfabric, libverbs, and any other transport layer including any optimizations of their implementation that will benefit application performance or workflow efficiency.

3.3.5 The Offeror shall provide a complete implementation of the OpenMP version 4.1 (or then current) standard including, if applicable, accelerator directives, as well as a supporting programming environment. The Offeror shall provide a detailed feature description of the OpenMP implementation(s) and describe any expected deviations from the OpenMP standard.

3.3.6 The Offeror shall provide a description of how OpenMP 3.1 applications will be compiled and executed on the proposed system.

3.3.7 The Offeror shall provide a description of any hardware or software features that enable OpenMP performance optimizations.

3.3.8 The Offeror shall list any PGAS languages and/or libraries that are supported (e.g. UPC, SHMEM, CAF, Global Arrays) and describe any system hardware and/or programming environment software provided that optimize any of the supported PGAS languages. The Offeror shall provide a mechanism to compile, run, and debug UPC applications.

3.3.9 The Offeror shall describe and list support for any emerging programming models such as asynchronous task/data models (e.g., Legion, STAPL, HPX, or OCR) and describe any system hardware and/or programming environment software provided that optimizes any of the supported models.

3.3.10 The Offeror shall describe the hardware and software environment support for:

- Fast thread synchronization of subsets of execution threads.
- Atomic add, fetch-and-add, multiply, bitwise operations, and compare-and-swap operations over integer, single-precision, and double-precision operands.
- Atomic compare-and-swap operations over 16-byte wide operands that comprise two double precision values or two memory pointer operands.
- Fast context switching or task-switching.
- Fast task spawning for unique and identical task with data dependencies.
- Support for active messages.

3.3.11 The Offeror shall describe in detail all programming APIs, languages, compliers and compiler extensions, etc. other than MPI and OpenMP (e.g. OpenACC, CUDA, OpenCL, etc.) that will be supported. It is desirable that instances of all programming models provided be interoperable and efficient when used within a single process or single job running on the same compute node.

3.3.12 The Offeror shall support the languages C, C++ (including complete C++11/14/17), Fortran 77, Fortran 90, and Fortran 2008 programming languages. Providing multiple compilation environments is highly desirable. The Offeror shall describe any limitations that can be expected in meeting full C++17 support based on current expectations.

3.3.13 The Offeror shall provide a Python implementation that will run on the compute partition with optimized MPI4Py, NumPy, and SciPy libraries.

3.3.14 The Offeror shall provide a programming toolchain(s) that enables runtime coexistence of threading in C, C++, and Fortran, from within applications and any supporting libraries using the same toolchain.

3.3.15 The Offeror shall provide C++ compiler(s) that can successfully build the Boost C++ library, http://www.boost.org. The Offeror shall support the most recent stable version of Boost.

3.3.16 The Offeror shall provide optimized versions of libm, libgsl, BLAS levels 1, 2 and 3, LAPACK, ScaLAPACK, HDF5, NetCDF, and FFTW. It is desirable for these to efficiently interoperate with applications that utilize OpenMP. The Offeror shall additionally describe all other optimized libraries that will be supported, including a description of the interoperability of these libraries with the programming environments proposed.

3.3.17 The Offeror shall provide a mechanism that enables control of task and memory placement within a node for efficient performance. The Offeror shall provide a detailed description of controls provided and any limitations that may exist.

3.3.18 The Offeror shall provide a comprehensive software development environment with configuration and source code management tools. On heterogeneous systems, a mechanism (e.g., an upgraded autoconf) shall be provided to create configure scripts to build cross-compiled applications on login nodes.

3.3.19 The Offeror shall provide an interactive parallel debugger with an X11-based graphical user interface. The debugger shall provide a single point of control that can debug applications in all supported languages using all granularities of parallelism (e.g. MPI+X) and programming environments provided and scale up to 25% of the proposed platform.

3.3.20 The Offeror shall provide a suite of tools for detailed performance analysis and profiling of user applications. At least one tool shall support all granularities of parallelism in mixed MPI+OpenMP programs and any additional programming models supported on the proposed system. The tool suite must provide the ability to support multi-node integrated profiling of on-node parallelism and communication performance analysis. The Offeror shall describe all proposed tools and the scalability limitations of each. The Offeror shall describe tools for measuring I/O behavior of user applications.

3.3.21 The Offeror shall provide event-tracing tools. Event tracing of interest includes: message-passing event tracing, I/O event tracing, floating point exception tracing, and message-passing profiling. The event-tracing tool API shall provide functions to activate and deactivate event monitoring during execution from within a process.

3.3.22 The Offeror shall provide single- and multi-node stack-tracing tools. The tool set shall include a source-level stack trace back, including an API that allows a running process or thread to query its current stack trace.

3.3.23 The Offeror shall provide tools to assist the programmer in introducing limited levels of parallelism and data structure refactoring to codes using any proposed programming models and languages. Tool(s) shall additionally be provided to assist application developers in the design and placement of the data structures with the goal of optimizing data movement/placement for the classes of memory proposed in the system.

3.3.24 The Offeror shall provide software licenses to enable the following number of simulataneous users on the system:

	Crossroads	NERSC-9
Compiler	20	100
Debugger	20	20

3.4 Platform Storage

Platform storage is certain to be one of the advanced technology areas included in any platform delivered in this timeframe. The APEX laboratories anticipate these emerging technologies will enable new usage models. With this in mind, an accompanying whitepaper, "APEX Workflows," is provided that describes how application teams use HPC resources today to advance scientific goals. The whitepaper is designed to provide a framework for reasoning about the optimal solution to these challenges. The whitepaper is intended to help an Offeror develop a storage architecture response that accelerates the science workflows while minimizing the total number of storage system tiers. The Crossroads/NERSC-9 workflows document can be found on the APEX website.

3.4.1 The Offeror shall provide a storage system capable of retaining all application input, output, and working data for 12 weeks (84 days), estimated at a minimum of 36% of baseline system memory per day.

3.4.2 The Offeror shall provide a storage system with an appropriate durability or a maintenance plan such that the storage system is capable of absorbing approximately four times baseline system memory per day for the life of the proposed system.

3.4.3 The Offeror shall describe how the proposed system provides sufficient bandwidth to support a JMTTI/Delta-Ckpt ratio of greater than 200 (where Delta-Ckpt is less than 7.2 minutes).

3.4.4 The Offeror shall describe the projected characteristics of all integrated storage devices, including but not limited to:

- Usable capacity, access latencies, storage interfaces (e.g. NVMe, PCIe), expected lifetime (warranty period, MTTF, total writes, etc.), and media and device error rates
- Relevant software/firmware features
- Compression technologies used by the storage devices, the resources used to implement the compression/decompression algorithms, the expected compression rates, and all compression/decompression-related performance impacts

3.4.5 The Offeror shall describe all available interfaces to storage, including but not limited to:

- POSIX
- APIs
- Exceptions to POSIX compliance.
- Time to consistency and any potential delays for reliable data consumption.
- Any special requirements for users to achieve performance and/or consistent data.

3.4.6 The Offeror shall describe the reliability characteristics of the proposed storage system, including but not limited to:

- Any single point of failure for all proposed storage tiers (note any component failure that will lead to temporary or permanent loss of data availability).
- Mean time to data loss for each storage tier provided.

- Enumerate storage tiers that are designed to be less reliable or do not use data protection techniques (e.g., replication, erasure coding).
- The magnitudes and duration of performance and reliability degradation brought about by a single or multiple component failures for each reliable storage tier.
- Vendor supplied mechanisms to ensure data integrity in each storage tier (e.g., data scrubbing processes, background checksum verification, etc.).
- Enumerate any storage system failures that potentially impact scheduled or currently executing jobs that impact the storage system performance and/or availability.
- Login or interactive nodes access to storage when the compute nodes are unavailable.

3.4.7 The Offeror shall describe system features for storage tier management designed to accelerate workflows, including but not limited to:

- Mechanisms for migrating data between storage tiers, including manual, scheduled, and/or automatic data migration to include rebalancing, draining, or rewriting data across devices within a tier.
- How storage will be instantiated with each job if it needs to be, and how storage may be persisted across jobs.
- The capabilities provided with the storage system to define per-user policies and automate data movement between different tiers of storage or different storage systems (e.g., archives).
- The ability to serialize namespaces no longer in use (e.g., snapshots).
- The ability to restore namespaces needed for a scheduled job that is not currently available.
- The ability to integrate with or act as a site-wide scheduling resource.
- A mechanism to incrementally add capacity and bandwidth to a particular tier of the storage system without requiring a tier-wide outage.
- Capabilities to manage or interface the proposed storage system with other storage systems or archives (e.g., fast storage layers or HPSS).

3.4.8 The Offeror shall describe software features that allow users to optimize I/O for the proposed workflows, including but not limited to:

- Batch data movement capabilities, especially when data resides on multiple tiers of storage.
- Methods for users to create and manage storage allocations.
- Any ability to directly write to or read from a tier not directly (logically) adjacent to the compute resources.
- Locality-aware job/data scheduling.
- I/O utilization for reservations.
- Features to prevent data duplication on more than one storage tier.

- Methods for users to exploit any enhanced performance of relaxed consistency.
- Methods for enabling user-defined metadata with the proposed storage solution.

3.4.9 The Offeror shall describe the method for walking the entire storage system's metadata, and describe any special capabilities that would mitigate user performance issues for daily full-system namespace walks; expect at least 1 billion objects.

3.4.10 The Offeror shall describe any capabilities to comprehensively collect storage system usage data (in a scalable way), including but not limited to:

- Per client metrics and frequency of collection, including but not limited to: the number of bytes read or written, number of read or write invocations, client cache statistics, and metadata statistics such as number of opens, closes, creates, and other system calls of relevance to the performance of the storage system.
- Job level metrics, such as the number of sessions each job initiates with each storage tier, session duration, total data transmitted (separated as reads and writes) during the session, and the number of total storage system invocations made during the session.
- Storage tier metrics and frequency of collection, such as the number of bytes read, number of bytes written, number of read invocations, number of write invocations, bytes deleted/purged, number of I/O sessions established, and periods of outage/unavailability.
- Job level metrics describing usage of a tiered storage hierarchy, such as how long files are resident in each tier, hit rate of file pages in each tier (i.e., whether pages are actually read and how many times data is re-read), fraction of data moved between tiers because of a) explicit programmer control and b) transparent caching, and time interval between accesses to the same file (e.g., how long until an analysis program reads a simulation generated output file).

3.4.11 The Offeror shall describe a method for providing access to the storage system from other systems at the facility. In the case of tiered storage, at least one tier must satisfy this requirement.

3.4.12 The Offeror shall describe the capability for storage tiers to be repaired, serviced, and incrementally patched/upgraded while running different versions of software or firmware without requiring a storage tier-wide outage. The Offeror shall describe the level of performance degradation, if any, experienced during the repair or service interval.

3.4.13 The Offerer shall specify the minimum number of compute nodes required to read and write the following data sets from/to the storage system:

- A 1 TB data set of 20 GB files in 2 seconds.
- A 1 PB data set of 32 MB files in 1 hour.

3.5 Application Performance

Assuring that real applications perform well on both the Crossroads and NERSC-9 platforms is key for their success. Because the full applications are large, often with millions of lines of code, and in some cases are export controlled, a suite of benchmarks have been developed for RFP response evaluation and system acceptance. The benchmark codes are representative of the workloads of the APEX laboratories but often smaller than the full applications.

The performance of the benchmarks will be evaluated as part of both the RFP response and platform acceptance. Final benchmark acceptance performance targets will be negotiated after a final system configuration is defined. All performance tests must continue to meet negotiated acceptance criteria throughout the lifetime of the proposed system.

Platform acceptance for Crossroads will also include an ASC Simulation Code Suite comprised of at least two (2) but no more than four (4) ASC applications from the three NNSA laboratories, Sandia, Los Alamos and Lawrence Livermore.

The Crossroads/NERSC-9 benchmarks, information regarding the Crossroads acceptance codes, and supplemental materials can be found on the APEX website.

3.5.1 The Offeror shall provide responses to the benchmarks (SNAP, PENNANT, HPCG, MiniPIC, UMT, MILC, MiniDFT, GTC, and Meraculous) provided on the *Crossroads/NERSC-9 benchmarks* link on the APEX website. All modifications or new variants of the benchmarks (including makefiles, build scripts, and environment variables) are to be supplied in the Offeror's response.

- The results of all problem sizes (baseline and optimized) shall be provided in the Offeror's Scalable System Improvement (SSI) spreadsheets. SSI is the calculation used for measuring improvement and is documented on the APEX website along with the SSI spreadsheets. If predicted or extrapolated results are provided, the methodology used to derive them should be documented.
- The Offeror shall provide licenses for the proposed system for all compilers, libraries, and runtimes used to achieve benchmark performance.

3.5.2 The Offeror shall provide performance results for the proposed system that may be benchmarked, predicted, and/or extrapolated for the baseline MPI+OpenMP (or UPC for Meraculous) variants of the benchmarks. The Offeror may modify the benchmarks to include extra OpenMP pragmas as required, but the benchmark must remain a standard-compliant program that maintains existing output subject to the validation criteria described in the benchmark run rules.

3.5.3 The Offeror shall optionally provide performance results from an Offeror optimized variant of the benchmarks. The Offeror may modify the benchmarks, including the algorithm and/or programming model used to demonstrate high system performance. If algorithmic changes are made, the Offeror should provide an explanation of why the results may deviate from validation criteria described in the benchmark run rules.

3.5.4 For the Crossroads system only: in addition to the *Crossroads/NERSC-9 benchmarks*, an ASC Simulation Code Suite representing the three NNSA laboratories will be used to judge performance at time of acceptance. The Crossroads system shall achieve a minimum of at least 6 times (6X) improvement over the ASC Trinity platform (Knights Landing partition) for each code, measured using SSI. The Offeror shall specify a baseline performance greater than or equal to 6X at time of response. Final acceptance performance targets will be negotiated after a final system configuration is defined. Information regarding ASC Simulation Code Suite run rules and acceptance can be found on the APEX website. Source code will be provided to the Offeror but will require compliance with export control laws and no cost licensing agreements.

3.5.5 The Offeror shall report or project the number of cores necessary to saturate the available node baseline memory bandwidth as measured by the Crossroads/NERSC-9 memory bandwidth benchmark found on the APEX website.

- If the node contains heterogeneous cores, the Offeror shall report the number of cores of each architecture necessary to saturate the available baseline memory bandwidth.
- If multiple tiers of memory are available, the Offeror shall report the above for every functional combination of core architecture and baseline or extended memory tier.

3.5.6 The Offeror shall report or project the sustained dense matrix multiplication performance on each type of processor core (individually and/or in parallel) of the proposed node architecture as measured by the Crossroads/NERSC-9 multithreaded DGEMM benchmark found on the APEX website.

- The Offeror shall describe the percentage of theoretical double-precision (64-bit) computational peak, which the benchmark GFLOP/s rate achieves for each type of compute core/unit in the response, and describe how this is calculated.

3.5.7 The Offeror shall report, or project, the MPI two-sided message rate of the node under the following conditions measured by the communication benchmark specified on the APEX website:

- Using a single MPI rank per node with MPI_THREAD_SINGLE.
- Using two, four, and eight MPI ranks per node with MPI_THREAD_SINGLE.
- Using one, two, four, and eight MPI ranks per node and multiple threads per rank with MPI_THREAD_MULTIPLE.

The Offeror may additionally choose to report on other configurations.

3.5.8 The Offeror shall report, or project, the MPI one-sided message rate of the node for all passive synchronization RMA methods with both pre-allocated and dynamic memory windows under the following conditions measured by the communication benchmark specified on the APEX website using:

- A single MPI rank per node with MPI_THREAD_SINGLE.
- Two, four, and eight MPI ranks per node with MPI_THREAD_SINGLE.
- One, two, four, and eight MPI ranks per node and multiple threads per rank with MPI_THREAD_MULTIPLE.

The Offeror may additionally choose to report on other configurations.

3.5.9 The Offeror shall report, or project, the time to perform the following collective operations for full, half, and quarter machine size and report on core occupancy during the operations measured by the communication benchmark specified on the APEX website for:

- An 8 byte MPI_Allreduce operation.
- An 8 byte per rank MPI_Allgather operation.

3.5.10 The Offeror shall report, or project, the minimum and maximum off-node latency for MPI two-sided messages using the following threading modes measured by the communication benchmark specified on the APEX website:

- MPI_THREAD_SINGLE with a single thread per rank.
- MPI_THREAD_MULTIPLE with two or more threads per rank.

3.5.11 The Offeror shall report, or project, the minimum and maximum off-node latency for MPI one-sided messages for all passive synchronization RMA methods with both pre-allocated and dynamic memory windows using the following threading modes measured by the communication benchmark specified on the APEX website:

- MPI_THREAD_SINGLE with a single thread per rank.
- MPI_THREAD_MULTIPLE with two or more threads per rank.

3.5.12 The Offeror shall provide an efficient implementation of MPI_THREAD_MULTIPLE. Bandwidth, latency, and message throughput measurements using the MPI_THREAD_MULTIPLE thread support level shall have no more than a 10% performance degradation when compared to using the MPI_THREAD_SINGLE support level as measured by the communication benchmark specified on the APEX website.

3.5.13 The Offeror shall report, or project, the maximum I/O bandwidths as measured by the IOR benchmark specified on the APEX website.

3.5.14 The Offeror shall report, or project, the metadata rates as measured by the MDTEST benchmark specified on the APEX website.

3.5.15 The Offeror shall be required at time of acceptance to meet specified targets for acceptance benchmarks, and mission codes for Crossroads, listed on the APEX website.

3.5.16 The Offeror shall describe how the proposed system may be configured to support a high rate and bandwidth of TCP/IP connections to external services both from compute nodes and directly to and from the storage system, including:

- Compute node external access shall allow all nodes to each initiate 1 connection concurrently within a 1 second window.
- Transfer of data over the external network to and from the compute nodes and the storage system at 100 GB/s per direction of a 1 TB dataset comprised of 20 GB files in 10 seconds.

3.6 Resilience, Reliability, and Availability

The ability to achieve the APEX mission goals hinges on the productivity of users of the platforms. Platform availability is therefore essential and requires system-wide focus to achieve a resilient, reliable, and available platform. For each metric specified below, the Offeror must describe how they arrived at their estimates.

3.6.1 Failure of the system management and/or RAS system(s) shall not cause a system or job interrupt. This requirement does not apply to a RAS system feature, which automatically shuts down the system for safety reasons, such as an overheating condition.

3.6.2 The minimum System Mean Time Between Interrupt (SMTBI) shall be greater than 720 hours.

3.6.3 The minimum Job Mean Time To Interrupt (JMTTI) shall be greater than 24 hours. Automatic restarts do not mitigate a job interrupt for this metric.

3.6.4 The ratio of JMTTI/Delta-Ckpt shall be greater than 200. This metric is a measure of the system's ability to make progress over a long period of time and corresponds to an efficiency of approximately 90%. If, for example, the JMTTI requirement is not met, the target JMTTI/Delta-Ckpt ratio ensures this minimum level of efficiency.

3.6.5 An immediate re-launch of an interrupted job shall not require a complete resource reallocation. If a job is interrupted, there shall be a mechanism that allows re-launch of the application using the same allocation of resource (e.g., compute nodes) that it had before the interrupt or an augmented allocation when part of the original allocation experiences a hard failure.

3.6.6 A complete system initialization shall take no more than 30 minutes. The Offeror shall describe the full system initialization sequence and timings. System initialization is defined to be the time to bring 99% of the compute resource and 100% of any service resource to the point where a job can be successfully launched.

3.6.7 The system shall achieve 99% scheduled system availability. System availability is defined in the glossary.

3.6.8 The Offeror shall describe the resilience, reliability, and availability mechanisms and capabilities of the proposed system including, but not limited to:

- Any condition or event that can potentially cause a job interrupt.
- Resiliency features to achieve the availability targets.
- Single points of failure (hardware or software), and the potential effect on running applications and system availability.
- How a job maintains its resource allocation and is able to relaunch an application after an interrupt.
- A system-level mechanism to collect failure data for each kind of component.

3.7 Application Transition Support and Early Access to APEX Technologies

The Crossroads and NERSC-9 platforms will include numerous pre-exascale technologies. The Offeror should include in their response a plan to effectively utilize these technologies and assist in transitioning the mission workflows to the proposed platforms. For the Crossroads platform only, the Offeror shall support efforts to transition the Advanced Technology Development Mitigation (ATDM) codes to the proposed platforms. ATDM codes are currently being developed by the three NNSA weapons laboratories, Sandia, Los Alamos, and Lawrence Livermore. These codes may require compliance with export control laws and no cost licensing agreements. Information about the ATDM program can be found on the NNSA website.

3.7.1 The Offeror shall propose a vehicle (e.g., a Center of Excellence) for supporting the successful demonstration of the application performance requirements and the transition of key applications to the Crossroads and NERSC-9 systems. Support will be required from the successful Offeror and all of its key advanced technology providers (e.g., processor vendors, integrators, etc). Activities will require the support of experts in the areas of application porting and performance optimization in the form of staff training, general user training, and deep-dive interactions with a set of application code teams. Support is required for compilers to enable timely bug fixes as well as enable new functionality. Support is required from the date of subcontract execution through two (2) years after final acceptance.

3.7.2 The Offeror shall describe which of the proposed APEX hardware and software technologies (physical hardware, emulators, and/or simulators) , will be available for access before platform delivery and in what timeframe. The proposed technologies should provide value in advanced preparation for the delivery of the final APEX platform(s) for pre-platform-delivery application porting and performance assessment activities.

3.8 Target System Configuration

Table 2 Target System Configuration

	Crossroads	NERSC-9
Baseline Memory Capacity ***Excludes all levels of on-die-CPU cache***	> 3 PiB	> 3 PiB
Benchmark SSI increase over Edison system	> 20X	> 20X
Platform Storage	> 30X Baseline Memory	> 30X Baseline Memory
Wall Plate Power	< 20 MW	< 20 MW
Peak Power	< 18 MW	< 18 MW
Nominal Power	< 15 MW	< 15 MW
Idle Power	< 10% Wall Plate Power	< 10% Wall Plate Power
Job Mean Time To Interrupt (JMTTI) ***Calculated for a single job running in the entire system***	> 24 Hours	> 24 Hours
System Mean Time To Interrupt (SMTTI)	>720 Hours	> 720 Hours
Delta-Ckpt	< 7.2 minutes	< 7.2 minutes
JMTTI/Delta-Ckpt	> 200	> 200
System Availability	> 99%	> 99%

3.9 System Operations

System management should be an integral feature of the overall system and should provide the ability to effectively manage system resources with high utilization and throughput under a workload with a wide range of concurrencies. The goal is to provide system administrators, security officers, and user-support personnel with productive and efficient system configuration management capabilities and an enhanced diagnostic environment.

3.9.1 The Offeror shall deliver scalable integrated system management capabilities that provide human interfaces and APIs for system configuration and its ability to be automated, software management, change management, local site integration, and system configuration backup and recovery.

3.9.2 The Offeror shall provide a means for tracking and analyzing all software updates, software and hardware failures, and hardware replacements over the lifetime of the proposed system.

3.9.3 The Offeror shall provide the ability to perform rolling upgrades and rollbacks on a subset of the system while the balance of the proposed system remains in production operation. The Offeror shall describe the mechanisms, capabilities, and limitations of rolling upgrades and rollbacks. No more than half the system partition shall be required to be down for rolling upgrades and rollbacks.

3.9.4 The Offeror shall provide an efficient mechanism for reconfiguring and rebooting compute nodes. The Offeror shall describe in detail the compute node reboot mechanism, differentiating types of boots (warmboot vs. coldboot) required for different node features, as well as how the time required to reboot scales with the number of nodes being rebooted.

3.9.5 The Offeror will ensure that all monitoring data and logs captured by the proposed system are available to the system owner, and will support an open monitoring API to facilitate lossless, scalable sampling and data collection for monitored data. Any filtering that may need to occur will be at the option of the system manager. The Offeror will provide a sampling and connection framework that allows the system manager to configure independent alternative parallel data streams to be directed off the system to site-configurable consumers.

3.9.6 The Offeror shall collect and provide metrics and logs monitoring the status, health, and performance of the proposed system, including, but not limited to:

- Environmental measurement capabilities for all systems and peripherals and their sub-systems and supporting infrastructure, including power and energy consumption and control.
- Internal HSN performance counters, including measures of network congestion and network resource consumption.
- All levels of integrated and attached storage.
- The system as a whole, including hardware performance counters for metrics for all levels of integrated and attached storage.

3.9.7 The Offeror shall describe what vendor-provided tools are available for the collection, analysis, integration, and visualization of metrics and logs produced by the proposed system (e.g., peripherals, integrated and attached storage, and environmental data, including power and energy consumption).

3.9.8 The Offeror shall describe the system configuration management and diagnostic capabilities that addresses the following topics:

- Detailed description of the system management support.
- Any effect or overhead of software management tool components on the CPU or memory available on compute nodes.
- Release plan, with regression testing and validation for all system related software and security updates.
- Support for multiple simultaneous or alternative system software configurations, including estimated time and effort required to install both a major and a minor system software update.
- User activity tracking, such as audit logging and process accounting.
- Unrestricted privileged access to all hardware components delivered with the system.

3.10 Power and Energy

Power, energy, and temperature will be critical factors in how the APEX laboratories manage the proposed platforms in this time frame and must be an integral part of overall Systems Operations. The proposed solution must also be well integrated into other intersecting areas (e.g., facilities, resource management, runtime systems, and applications). The APEX laboratories expect a growing number of use cases in this area that will require a vertically integrated solution.

3.10.1 The Offeror shall describe all power, energy, and temperature measurement capabilities (platform, rack/cabinet, board, node, component, and sub-component level), including control and response times, sampling frequency, accuracy of the data, and timestamps of the data for individual points of measurement and control.

3.10.2 The Offeror shall describe all control capabilities provided to affect power or energy use (platform, rack/cabinet, board, node, component, and sub-component level).

3.10.3 The Offeror shall provide platform-level interfaces that enable measurement and dynamic control of power and energy relevant characteristics of the proposed system, including but not limited to:

- AC measurement capabilities at the platform or rack level.
- Platform-level minimum and maximum power settings (e.g., power caps).
- Platform-level power ramp up and down rate.

- Scalable collection and retention all measurement data such as:
- point-in-time power data.
- energy usage information.
- minimum and maximum power data.

3.10.4 The Offeror shall provide resource manager interfaces that enable measurement and dynamic control of power and energy relevant characteristics of the proposed system, including but not limited to:

- Job and node level minimum and maximum power settings.
- Job and node level power ramp up and down rate.
- Job and node level processor and/or core frequency control.
- Platform and job level profiling and forecasting.
    - e.g., prediction of hourly power averages >24 hours in advance with a 1 MW tolerance.

3.10.5 The Offeror shall provide application and runtime system interfaces that enable measurement and dynamic control of power and energy relevant characteristics of the proposed system including but not limited to:

- Node level minimum and maximum power settings.
- Node level processor and/or core frequency control.
- Node level application hints, such as:
    - application entering serial, parallel, computationally intense, I/O intense or communication intense phase.

3.10.6 The Offeror shall provide an integrated API for all levels of measurement and control of power relevant characteristics of the proposed system. It is preferable that the provided API complies with the High Performance Computing Power Application Programming Interface Specification (http://powerapi.sandia.gov).

3.10.7 The Offeror shall project the Wall Plate, Peak, Nominal, and Idle Power of the proposed system.

3.10.8 The Offeror shall describe any controls available to enforce or limit power usage below wall plate power and the reaction time of this mechanism (e.g., what duration and magnitude can power usage exceed the imposed limits).

3.10.9 The Offeror shall describe the status of the proposed platform when in an Idle State (describe all Idle States if multiple are available) and the time to transition from the Idle State (or each Idle State if there are multiple) to the start of job execution.

3.11 Facilities and Site Integration

3.11.1 The proposed system shall use 3-phase 480V AC. Other system infrastructure components (e.g., disks, switches, login nodes, and mechanical subsystems such as CDUs) must use either 3-phase 480V AC (strongly preferred), 3-phase 208V AC (second choice), or single-phase 120/240V AC (third choice). The total number of individual branch circuits and phase load imbalance should be minimized.

3.11.2 All equipment and power control hardware shall be Nationally Recognized Testing Laboratories (NRTL) certified and bear appropriate NRTL labels.

3.11.3 Every rack, network switch, interconnect switch, node, and disk enclosure shall be clearly labeled with a unique identifier visible from the front of the rack and/or the rear of the rack, as appropriate, when the rack door is open. These labels will be high quality so that they do not fall off, fade, disintegrate, or otherwise become unusable or unreadable during the lifetime of the proposed system. Nodes will be labeled from the rear with a unique serial number for inventory tracking. It is desirable that motherboards also have a unique serial number for inventory tracking. This serial number needs to be visible without having to disassemble the node, or else it must be able to be queried from the system management console.

3.11.4 The Offeror shall describe the features of the proposed system related to facilities and site integration, including:

- Description of the physical packaging of the proposed system, including dimensioned drawings of individual cabinets types and the floor layout of the entire system.
- Remote environmental monitoring capabilities of the system and how it would integrate into facility monitoring.
- Emergency shutdown capabilities.
- Detailed descriptions of power and cooling distributions throughout the system, including power consumption for all subsystems.
- Description of parasitic power losses within Offeror's equipment, such as fans, power supply conversion losses, power-factor effects, etc. For the computational and storage subsystems separately, give an estimate of the total power and parasitic power losses (whose difference should be power used by computational or storage components) at the minimum and maximum ITUE, which is defined as the ratio of total equipment power over power used by computational or storage components. Describe the conditions (e.g. "idle") at which the extrema occur.
- OS distributions or other client requirements to support off-platform access to the parallel file system (e.g. LANL File Transfer Agents).

Table 3 Crossroads and NERSC-9 Facility Requirements

	Crossroads	NERSC-9
Location	Los Alamos National Laboratory, Los Alamos, NM. The proposed system will be housed in the Strategic Computing Complex (SCC), Building 2327	National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA. The proposed system will be housed in Wang Hall, Building 59 (formerly known as the Computational Theory and Research Facility).
Altitude	7,500 feet	650 feet
Seismic	N/A	System to be placed on a seismic isolation floor. System cabinets should have an attachment mechanism that will enable them to be firmly attached to each other and the isolation floor. When secured via these attachments, the cabinets should be able to withstand seismic design accelerations per the California Building Code and LBNL Lateral Force Design Criteria policy in effect at the time of subcontract award. (The CBC currently specifies 0.49g but is expected to be updated in 2016.)
Water Cooling	The system must operate in conformance with ASHRAE Class W2 guidelines (dated 2011). The facility will provide operating water	Same

	Crossroads	NERSC-9
	temperature that nominally varies between 60-75°F, at up to 35PSI differential pressure at the system cabinets However, Offeror should note if the system is capable of operating at higher temperatures. Note: LANL facility will provide inlet water at a nominal 75°F. It may go to as low as 60°F based on facility and/or environmental factors. Total flow requirements may not exceed 9600GPM.	Note: NERSC facility will provide inlet water at a nominal 65°F. It may go as high as 75°F based on facility and/or environmental factors. Total flow requirements may not exceed 9600GPM.
Water Chemistry	The system must operate with facility water meeting basic ASHRAE water chemistry. Special chemistry water is not available in the main building loop and would require a separate tertiary loop provided with the system. If tertiary loops are included in the system, the Offeror shall describe their operation and maintenance, including coolant chemistry, pressures, and flow controls. All coolant loops within the system shall have reliable leak detection, temperature, and flow alarms, with automatic protection and	Same

	Crossroads	NERSC-9
	notification mechanisms.	
Air Cooling	The system must operate with supply air at 76°F or below, with a relative humidity from 30%-70%. The rate of airflow is between 800-1500 CFM/floor tile. No more than 3MW of heat shall be removed by air cooling.	The system must operate with supply air at 76°F or below, with a relative humidity from 30%-80%. The current facility can support up to 60K CFM of airflow, and remove 500KW of heat. Expansion is possible to 300K CFM and 1.5MW, but at added expense.
Maximum Power Rate of Change	The hourly average in platform power should not exceed the 2MW wide power band negotiated at least 2 hours in advance.	N/A
Power Quality	The system shall be resilient to incoming power fluctuations at least to the level guaranteed by the ITIC power quality curve.	Same
Floor	42” raised floor	48” raised floor
Ceiling	16 foot ceiling and an 18’ 6” ceiling plenum	17’10” ceiling however maximum cabinet height is 9’5”
Maximum Footprint	8000 square feet; 80 feet long and 100 feet deep.	64’x92’, or 5888 square feet (inclusive of compute, storage and service aisles). This area is itself surrounded by a minimum 4’ aisle that can be used in the system layout. It is preferred that cabinet rows run parallel to the short dimension.
Shipment Dimensions and	No restrictions.	For delivery, system components shall weigh

	Crossroads	NERSC-9
Weight		less than 7000 pounds and shall fit into an elevator whose door is 6ft 6in wide and 9ft 0 in high and whose depth is 8ft 3in. Clear internal width is 8ft 4 in.
Floor Loading	The average floor loading over the effective area shall be no more than 300 pounds per square foot. The effective area is the actual loading area plus at most a foot of surrounding fully unloaded area. A maximum limit of 300 pounds per square foot also applies to all loads during installation. The Offeror shall describe how the weight will be distributed over the footprint of the rack (point loads, line loads, or evenly distributed over the entire footprint). A point load applied on a one square inch area shall not exceed 1500 pounds. A dynamic load using a CISCA Wheel 1 size shall not exceed 1250 pounds (CISCA Wheel 2 – 1000 pounds).	The floor loading shall not exceed a uniform load of 500 pounds per square foot. Raised floor tiles are ASM FS400 with an isolated point load of 2000 pounds and a rolling load of 1200 pounds.
Cabling	All power cabling and water connections shall be below the access floor. All other cabling (e.g., system interconnect) should be above floor and integrated into the	Same

	Crossroads	NERSC-9
	system cabinetry. Under floor cables (if unavoidable) shall be plenum rated and comply with NEC 300.22 and NEC 645.5. All communications cables, wherever installed, shall be source/destination labeled at both ends. All communications cables and fibers over 10 meters in length and installed under the floor shall also have a unique serial number and dB loss data document (or equivalent) delivered at time of installation for each cable, if a method of measurement exists for cable type.	
External network interfaces supported by the site for connectivity requirements specified below	1Gb, 10Gb, 40Gb, 100Gb, IB	Same
External bandwidth on/off the system for general TCP/IP connectivity	> 100 GB/s per direction	Same
External bandwidth on/off the system for accessing the system's PFS	> 100 GB/s	Same
External bandwidth on/off the system for accessing external, site supplied file systems. E.g. GPFS, NFS	> 100 GB/s	Same

4 Non-Recurring Engineering

The APEX team expects to award a Non-Recurring Engineering (NRE) subcontract as part of each platform subcontract to the selected Offeror (one NRE subcontract per platform, awarded by the contracting organization). It is expected that Crossroads and NERSC personnel will collaborate in both NRE subcontracts. It is anticipated that the NRE subcontracts will be approximately 10%-15% of the combined Crossroads and NERSC-9 platform budgets. The Offeror is encouraged to provide proposals for areas of collaboration they feel provide substantial value to the Crossroads and NERSC-9 systems with the goals of:

- Increasing application performance.
- Increasing workflow performance.
- Increasing the resilience, and reliability of the system.

Proposed collaboration areas shall focus on topics that provide added value beyond planned roadmap activities. Proposals shall not focus on one-off point solutions or gaps created by their proposed design that should be otherwise provided as part of a vertically integrated solution. It is expected that NRE collaborations will have impact on both the Crossroads and NERSC-9 platforms and follow-on platforms procured by the U.S. Department of Energy's NNSA and Office of Science.

NRE topics of interest include, but are not limited to, the following:

- Development and optimization of hardware and software capabilities to increase the performance of MPI+OpenMP and future task-based asychronous programming models.
- Development and optimization of hardware and software capabilities to increase the performance of application workflows, including consideration of consistency requirements, data-migration needs, and system-wide resource management.
- Development of scalable system management capabilities to enhance the reliability, resilience, power, and energy usage of Crossroads/NERSC-9.

5 Options

The APEX team expects to have future requirements for system upgrades and/or additional quantities of components based on the configurations proposed in response to this solicitation. To address these potential requirements, the Offeror shall propose and separately price options for system upgrades and expansions as indicated in Section 5.1. The Offeror shall address any technical challenges foreseen with respect to scaling and any other production issues. Proposals should be as detailed as possible, and those that do not address all of the additional system options in a materially responsive manner will be downgraded.

5.1 Upgrades, Expansions and Additions

5.1.1 Upgrade, expand or procure additional system configurations by the following fractions of the proposed system as measured by the Sustained System Improvement (SSI) metric:

- 25%
- 50%
- 100%
- 200%

5.1.2 The Offeror shall propose a configuration or configurations which double the baseline memory capacity.

5.1.3 Upgrade, expand or procure additional storage system capacity (per tier if multiple tiers are present) in increments of 25% relative to proposed storage system.

5.2 Early Access Development System

To allow for early and/or accelerated development of applications or development of functionality required as a part of the statement of work, the Offeror shall propose options for early access development systems. These systems can be in support of the baseline requirements or any proposed options.

5.2.1 The Offeror shall propose an Early Access Development System. The primary purpose is to expose the application to the same programming environment as will be found on the final system. It is acceptable for the early access system to not use the final processor, node, or high-speed interconnect architectures. However, the programming and runtime environment must be sufficiently similar that a port to the final system is trivial. The early access system shall contain similar functionality of the final system, including file systems, but scaled down to the appropriate configuration. The Offeror shall propose an option for the following configurations based on the size of the final Crossroads/NERSC-9 systems.

- 2% of the compute partition.
- 5% of the compute partition.
- 10% of the compute partition.

5.2.2 If applicable, the Offeror shall propose development test bed systems that will reduce risk and aid the development of any advanced functionality that is exercised as a part of the statement of work. For example, any topics proposed for NRE.

5.3 Test Systems

The Offeror shall propose the following test systems. The systems shall contain all the functionality of the main system, including file systems, but scaled down to the appropriate configuration. Multiple test systems may be awarded.

5.3.1 The Offeror shall propose an Application Regression test system, which shall contain at least 200 compute nodes.

5.3.2 The Offeror shall propose a System Development test system, which shall contain at least 50 compute nodes.

5.4 On Site System and Application Software Analysts

5.4.1 The Offeror shall propose and separately price two (2) System Software Analysts and two (2) Applications Software Analysts for each site. For Crossroads, these positions require a DOE Q-clearance for access.

5.5 Maintenance and Support

The Offeror shall propose separately priced maintenance and support options with the following features:

5.5.1 Pricing and the Maintenance Period

The Offeror shall propose all technical services and maintenance options for a period of four (4) years from the date of acceptance of the system. Warranty shall be included in the 4 years. For example, if the system is accepted on April 1, 2021 and the Warranty is for one year, then the Warranty ends on March 30, 2022, and the maintenance period begins April 1, 2022 and ends on March 30, 2025.

While pricing is per year, the contracting organization may elect to pay in quarterly or monthly proportional installments rather than yearly. Contractor may terminate maintenance with at least 90 days notice. If maintenance is terminated mid-year, the maintenance costs shall be recalculated proportional to the time under maintenance, and any balance refunded.

5.5.2 Maintenance and Support Options

The Offeror shall propose each of the options below. The contracting organization may purchase or execute one of the Options or none of the Options at its discretion. Different maintenance options may be selected for the various test systems and final system. Each Option shall be separately priced. The Offeror may propose other Maintenance Solutions in addition to Options 1 and 2 below.

5.5.2.1 Option 1 – 7x24

The Offeror shall price Option 1 as full hardware and software support for all Offeror provided hardware components and software. The principal period of maintenance (PPM) shall be for 24 hours by 7 days a week with a four hour response to any request for service.

5.5.2.2 Option 2 – 5x9

The Offeror shall price Option 2 as full hardware and software support for all Offeror provided hardware components and software. The principal period of maintenance (PPM) shall be on a 9 hours by 5 days a week (exclusive of holidays observed by ACES or NERSC). The Offeror shall provide hardware maintenance training for ACES/NERSC staff so that staff are able to provide hardware support for all other times the Offeror is unable to provide hardware repair in a timely manner outside of the PPM. The Offeror shall supply hardware maintenance procedural documentation, training, and manuals necessary to support this effort.

All proposed maintenance solutions shall include the following features and meet all requirements of this section.

5.5.3 General Service Provisions

The successful Offeror shall be responsible for repair or replacement of any failing hardware component that it supplies and correction of defects in software that it provides as part of the system.

At its sole discretion, the contracting organization may request advance replacement of components which show a pattern of failures which reasonably indicates that future failures may occur in excess of reliability targets, or for which there is a systemic problem that prevents the contracting organization's effective use of the system.

Hardware failures due to environmental changes in facility power and cooling systems which can be reasonably anticipated (such as brown-outs, voltage-spikes or cooling system failures) are the responsibility of the successful Offeror.

The successful Offeror and Contractor shall work together on a "no fault" basis to diagnose and correct failures that affect operation of the system and that involve Contractor-supplied hardware or software. The successful Offeror shall provide assistance within the normal response hours, and is responsible for correcting any defects (if any) in the successful Offeror-supplied equipment or software.

5.5.4 Software and Firmware Update Service

The successful Offeror shall provide an update service for all software and firmware provided for the duration of the Warranty plus Maintenance period. This shall include new releases of software/firmware and software/firmware patches as required for for normal use. The successful Offeror shall integrate software fixes, revisions or upgraded versions in supplied software, including community software (e.g. Linux or Lustre), and make them available to each contracting organization within 12 months of their general availability. The successful Offeror shall provide prompt availability of patches for cybersecurity defects.

5.5.5 Call Service

The successful Offeror shall provide contact information for technical personnel with knowledge of the proposed equipment and software. These personnel shall be available for consultation by telephone and electronic mail with ACES/NERSC personnel. In the case of degraded performance, the successful Offeror's services shall be made readily available to develop strategies for improving performance, i.e. patches, workarounds.

5.5.6 On-site Parts Cache

The successful Offeror shall maintain a parts cache on-site at both the ACES and NERSC facilities. The parts cache shall be sized and provisioned sufficiently to support all normal repair actions for two weeks without the need for parts refresh. The initial sizing and provisioning of the cache shall be based on the successful Offeror's Mean Time Between Failure (MTBF) estimates for each FRU and each rack, and scaled based on the number of FRU's and racks delivered. The parts cache configuration will be periodically reviewed for quantities needed to satisfy this requirement, and adjusted if necessary, based on observed FRU or node failure rates. The parts cache will be resized, at the the successful Offeror's expense, should the on-site parts cache prove to be insufficient to sustain the actually observed FRU or node failure rates.

5.5.7 On-Site Node Cache

The successful Offeror shall also maintain an on-site spare node inventory of at least 1% of the total nodes in all of the system. These nodes shall be maintained and tested for hardware integrity and functionality utilizing the Hardware Support Cluster defined below if provided.

5.5.8 Deinstallation

The Offeror shall provide an option to deinstall, remove and recycle the system at end of life. Storage media shall be wiped or destroyed to the satisfaction of the contracting organization, or returned to the contracting organization at its request.

The following features and requirements are specific to responses for the ACES solutions.

5.5.9 Hardware Support Cluster (HSC)

The successful Offeror shall provide a Hardware Support Cluster. The HSC shall support the hot spare nodes and provide functions such as hardware burn-in, problem diagnosis, etc. The successful Offeror will supply sufficient racks, interconnect, networking, storage equipment and any associated hardware/software necessary to make the HSC a stand-alone system capable of running diagnostics on individual or clusters of HSC nodes. ACES will store and inventory the HSC and other on-site parts cache components.

5.5.10 DOE Q-Cleared Technical Service Personnel

The Crossroads systems will be installed in machine rooms and buildings located inside of the LANL and SNL security areas, which require a DOE Q-clearance for access. It will be possible to install the systems with the assistance of uncleared US citizens or L-cleared personnel, but the successful Offeror shall be required to arrange and to pay for appropriate 3rd party security escorts. The successful Offeror on-site support staff shall obtain the necessary clearances to perform their duties.

6 Delivery and Acceptance

Testing of the system shall proceed in three steps: pre-delivery, post-delivery, and acceptance. Each step is intended to validate the system and feeds into subsequent activities. Sample Acceptance Test plans shall be provided as part of the Request for Proposal.

6.1 Pre-delivery Testing

The APEX team and vendor staff shall perform pre-delivery testing at the factory on the hardware to be delivered. Any limitations for performing the pre-delivery testing need to be identified, including scale and licensing limitations (if any). During pre-delivery testing, the successful Offeror shall:

- Demonstrate RAS capabilities and robustness using simple fault injection techniques, such as disconnecting cables, powering down subsystems, or installing known bad parts.
- Demonstrate functional capabilities on each segment of the system built, including the capacity to build applications, schedule jobs, and run them using a customer-provided testing framework. The root cause of application failure must be identified prior to system shipping.
- Provide a file system sufficiently provisioned to support the suite of tests.
- Provide onsite and remote access to the APEX team to monitor testing and analyze results.

- Instill confidence in the ability to conform to the statement of work.

6.2 Site Integration and Post-delivery Testing

The APEX team and vendor staff shall perform site integration and post-delivery testing on the fully delivered system. Limitations and/or special requirements may exist for vendor access to the onsite system.

- During post-delivery testing, the pre-delivery tests shall be run on the full system installation.
- Where applicable, tests shall be run at full scale.

6.3 Acceptance Testing

The APEX team and vendor staff shall perform onsite acceptance testing on the fully installed system. Limitations and/or special requirements may exist for vendor access to the onsite system.

6.3.1 The vendor shall demonstrate that the delivered system conforms to the subcontract's Statement of Work. A sample test plan will be provided as part of the Request for Proposal.

7 Risk and Project Management

The Offeror's proposal shall:

7.1.1 Provide a risk management strategy for the proposed system in case of technology problems or scheduling delays that affect delivery of the system or achievement of performance targets in the proposed timeframe. Describe the impact of substitute technologies (if any) on the overall architecture and performance of the proposed system in particular adressing the four technology areas listed below:

- Processor
- Memory
- High-speed interconnect
- Platform storage

7.1.2 Identify any other high-risk areas and accompanying mitigation strategies for the proposed system.

7.1.3 Provide a clear plan for effectively responding to software and hardware defects and system outages at each severity level and document how problems or defects will be escalated.

7.1.4 Provide a roadmap showing how the response to this procurement aligns with their plans for exascale computing.

7.1.5 Discuss additional capabilities, including the Offeror's:

- Ability to produce and maintain the proposed system for the life of the platform
- Ability to achieve specific quality assurance, reliability, availability and serviceability goals
- In-house testing and problem diagnosis capability, including hardware resources at appropriate scale

7.1.6 Project management specifics for the APEX team will be detailed as part of the Request for Proposal document.

8 Documentation and Training

The Offeror shall provide documentation and training to effectively operate, configure, maintain, and use the proposed solution to the APEX team and users of the Crossroads and NERSC-9 systems. The APEX team may, at their option, make audio and video recordings of presentations from vendor's speakers at public events targeted at the APEX user communities (e.g., user training events, collaborative application events, best practices discussions, etc.). The vendor will grant the APEX team user and distribution rights of vendor-provided documentation, session materials, and recorded media to be shared with other DOE Labs' staff and all authorized users and support staff for Crossroads and NERSC-9.

8.1 Documentation

8.1.1 The Offeror shall provide documentation for each delivered system describing the configuration, interconnect topology, labeling schema, hardware layout, etc. of the system as deployed before the commencement of system acceptance testing.

8.1.2 The Offeror shall supply and support system- and user-level documentation for all components before the delivery of the system. Upon request by the laboratories, the Offeror shall supply additional documentation necessary for operation and maintenance of the system. All user-level documentation shall be publically avaliable.

8.1.3 All documentation shall be distributed and updated electronically and in a timely manner. For example, changes to the system shall be accompanied by relevant documentation. Documentation of changes and fixes may be distributed electronically in the form of release notes. Reference manuals may be updated later, but effort should be made to keep all documentation current.

8.2 Training

8.2.1 The Offeror shall provide the following types of training at facilities specified by ACES or NERSC:

	Number of Classes	
Class Type	ACES	NERSC
System Operations and Advanced Administration	2	2
User Programming	3	3

8.2.2 The Offeror shall describe all proposed training and documentation relevant to the proposed solutions utilizing the following methods:

- Classroom training
- Onsite training
- Online documentation
- Online training

9 References

APEX schedule and high-level information can be found at the primary APEX website http://apex.lanl.gov.

Crossroads/NERSC-9 benchmarks and workflows whitepaper can be found at the APEX Benchmark and Workflows website https://www.nersc.gov/research-and-development/apex/apex-benchmarks-and-workflows.

High Performance Computing Power Application Programming Interface Specification http://powerapi.sandia.gov.

Definitions and Glossary

Baseline Memory: High performance memory technologies such as DDR-DRAM, HBM, and HMC, for example, that may be included in the proposed platforms memory capacity requirement. It does not include memory associated with caches.

Coefficient of Variation: The ratio of the standard deviation to the mean.

Delta-Ckpt: The time to checkpoint 80% of aggregate memory of the system to persistent storage. For example, if the aggregate memory of the compute partition is 3 PiB, Delta-Ckpt is the time to checkpoint 2.4 PiB. Rationale: This will provide a checkpoint efficiency of about 90% for full system jobs.

Ejection Bandwidth: Bandwidth leaving the node (i.e., NIC to router).

Full Scale: All of the compute nodes in the system. This may or may not include all available compute resources on a node, depending on the use case.

Idle Power: The projected power consumed on the proposed platform when the system is in an **Idle State**.

Idle State: A state when the system is prepared to but not currently executing jobs. There may be multiple idle states.

Injection Bandwidth: Bandwidth entering the node (i.e., router to NIC).

Job Interrupt: Any system event that causes a job to unintentionally terminate.

Job Mean Time to Interrupt (JMTTI): Average time between job interrupts over a given time interval on the full scale of the proposed platform. Automatic restarts do not mitigate a job interrupt for this metric.

JMTTI/Delta-Ckpt: Ratio of the JMTTI to Delta-Ckpt, which provides a measure of how much useful work can be achieved on the system.

Nominal Power: The projected power consumed on the proposed platform by the APEX workflows (e.g., a combination of the APEX benchmark codes running large problems on the entire platform).

Peak Power: The projected power consumed by an application that utilizes the maximum achievable power consumption such as DGEMM.

Rolling Upgrades/Rolling Rollbacks: A rolling upgrade or a rollback is defined as changing the operating software or firmware of a system component in such a way that the change does not require synchronization across the entire system. Rolling upgrades and rollbacks are designed to be performed with those parts of the system that are not being worked on remaining in full operational capacity.

System Interrupt: Any system event, or accumulation of system events over time, resulting in more than 1% of the compute resource being unavailable at any given time. Loss of access to any dependent subsystem (e.g., storage-system or service partition resource) will also incur a system interrupt.

System Mean Time Between Interrupt (SMTBI): Average time between system interrupts over a given time interval.

System Availability: ((time in period – time unavailable due to outages in period)/(time in period – time unavailable due to scheduled outages in period)) * 100

System Initialization: The time to bring 99% of the compute resource and 100% of any service resource to the point where a job can be successfully launched.

Wall Plate (Name Plate) Power: The maximum theoretical power the proposed platform could consume. This is a design limit, likely not achievable in operation.